1 Overview

Deriving scientific information from large, complex datasets can motivate large, complex statistical models; below we will give examples from our work on problems in astronomy, genomics, phylogenetics, econometrics, internet advertising, and ecology. As models grow in complexity, the need to interrogate their assumptions, to propagate uncertainty amongst their components, and to perform non-parametric checks on their data sensitivity grows commensurately, but so does the computational cost of doing so using traditional statistical methods. Many classical procedures designed to address these concerns, such as Markov Chain Monte Carlo (MCMC), cross validation, or re-estimating a model under a range of modeling assumptions, can be prohibitively expensive in many modern problems.

To help fill this gap, my research focuses on applications of *sensitivity analysis*, applied not merely to assess sensitivity of modeling assumptions (though I do pursue this traditional role as well), but also to assess freqeuntist uncertainty and propagate uncertainty in Bayesian procedures. At its core, all my methodological work is based on using Taylor series approximations, constructed only from properties of a single model fit, computed using either optimization or MCMC, to extrapolate to alternatives without expensive re-estimation. Though venerable and conceptually simple, I show that sensitivity analysis has wide-ranging and fundamental applications in modern, computationally intensive statistical problems, and that certain theoretical ideas concerning the role of differential approximations in statistical analysis deserve re-examination in the age of big data and big computing.

2 Projects

2.1 Prior sensitivity in Bayesian analysis

Bayesian analysis allows analysts to reason coherently about unknown parameters, but only if the user specifies a complete generating process for the parameters and data, including both prior distributions for the parameters and precise likelihoods for the data. Often aspects of this model are at best a considered simplifcation, and at worst chosen only for computational convenience. It is critical to ask whether the analysis would have changed substantively had different modeling choices been made.

Bayesian nonparametrics. A commonly asked question in unsupervised clustering is how many distinct clusters are present in a dataset. Discrete Bayesian nonparametrics allows the question to be addressed using Bayesian inference, but one must specify a prior on how distinct clusters are generated. A particularly common choice is the stick-breaking representation of a Dirichlet process prior, a mathematical abstraction arguably better justified by its mathematical convenience than its realism. The prior must be specified in terms of random stick lengths to be broken off successively, the lengths of the sticks determining the a priori cluster sizes. The standard approach is to model the stick lengths with $Beta(1, \alpha)$ distribution, the α being a scalar tuning parameters. Other classes of distributions are possible in principle but hardly ever considered in practice, in part because the Beta distribution enjoys some computational conveniences.

In CITE, we provide sensitivity measures that allow the user to explore alternative stick breaking distributions from a single fit using the standard and convenient Beta prior. We linearly approximating the dependence of the optimum on the functional form of a parameterized class of priors. A natural parameterized class is the set of $Beta(1, \alpha)$ distributions (parameterized by α), but we also consider arbitrarily functional perturbations. In current work in progress, we evaluate the worst-case perturbations. On a real-world clustering problem, a human genome dataset, we find that the number of distinct inferred populations is in fact quite sensitive to the prior.

Partial pooling in meta-analysis. A popular form of meta-analysis is to place a hierarchial model on a set of related experimental results, which both "shrinks" the individual estimates towards a common mean, potentially decreasing mean squared error, and allowing direct estimation of the average effect and diversity of effects. These advantages come at the cost of positing a precise generative process for the effects in question, however, and it is reasonable to interrogate whether the estimation procedure is robust to varaibility in these effects. In CITE, we apply sensitivity analysis to a published meta-analysis of the effectiveness of microcredit interventions in seven developing countries. We find that the conclusion are highly sensitive to the assumed covaraince structure between the base level of business profitability and the microcredit effect, a covariance which is a priori difficult to ascertain. In this way, we were able to easily diagnose a conceptual problem in a model which was time-consuming to fit.

Hyperparameter sensitivity for MCMC. A classical result in Bayesian sensitivity analysis states that derivatives of posterior expectations take the form of particular posterior covariances. The resulting sensitivities can be automatically computed in a black-box manner when the posterior is implemented in software that supports automatic differentiation, such as the popular Hamiltonian Monte Carlo sampler and modleing language, Stan. I have written an R package CITE that allows Stan users to specify a "hyperparameters" modeling block, from which one can automatically compute hyperparameter sensitivity from a single MCMC run with no additional computation. I apply these principles in a related work on frequentist variance below.

2.2 Data sensitivity: cross validation and frequentist variance

Frequentist variability is ultimately concerned with the value of an estimation procedure if the data were different than that observed. A classical mainfestiation of this idea is the nonparametric bootstrap, which estimates frequentist variability by evaluating a particular estimator at pseudo-datasets with observations draw with replacement from the observed dataset. Similarly, cross-validation (CV) in its various forms evaluates how a statistical procedure performs on data that were not included as part of estimation, and can be thought of as a non-parametric estimator of the bias induced by evaluating a loss function using the same data that were used to fit a model.

Both the bootstrap and CV require re-fitting a model with new, nearby datasets multiple times. When the model is differentiable, and model re-fitting is expensive, it can be advantageous to approximate the effect of re-fitting rather than perform actual re-fitting. One way of doing so it so is to perform a Taylor series expansion of the estimator, as a funciton of the empirical distribution, around the original empirical distribution. This is the core concept behind the related classical tools known as the "infinitesimal jackknife," "von-Mises Expansion," and "empirical influence function," though until recently these differnetial approximations were used most prominently to facilitate theoretical analysis.

Accuracy bounds for leave-k-out CV. We, and several other authors, obsevered that these differnetial methods could speed up the evaluation of cross validation in large machine learning models which are expensive to re-fit. In CITE, we bridged the gap from some of the classical literature, providing finite-sample accuracy bounds for approximate leave-k-out cross validation, even when the derivatives of the objective function are unbounded.

Higher orders, k-fold CV, and the bootstrap A follow-on work in progress (CITE) expands the results to higher-order expansions and to larger perturbations, including k-fold CV and the bootstrap. They key to all this work is a set complexity condition, in light of which it is clear that one can provide accuracy bounds uniformly over small perturbations, and over randomly-chosen large perturbations, but not for uniform bounds over large perturbations. Because we show that the linear approximation approaches the bootstrap closer than the bootstrap approaches the truth, our work should allow for practical differential approximations to prohibitively expensive procedures such as the bootstrap-after-the-bootstrap.

Frequentist properties of Bayesian posteriors By combining the above approach to frequentist variance with the MCMC-based measures of sensitivity, we are able to derive the Bayesian infinitesimal jackknife (IJ), which can be used to compute the frequentist variability of Bayeisan posterior means without bootstrapping or computing a maximum a-posteriori (MAP) estimate. Such frequentist variances are particularly important when there is a possibility of model misspecification (in which case the Bayesian posterior variance is not particularly meaningful), or when the data comes from a random sample, the variability of which is meaningful in its own right. In CITE, we prove the consistency of the Bayesian IJ and show its accuracy as an approximation to the bootstrap for a larger number of examples.

Adversarial sensitivity for M-estimators. In the social sciences, it is common to run randomized trials on a particular population to esitmate an effect that we hope generalizes to different populations. Classical standard errors measure the variability that one would expect from random sampling from the same distribution as that that gave rise to the observed data. However, in order to generalize to radically different contexts, one might expect the conclusions to be robust to more adversarial perturbations. In CITE, co-authors an I ask whether some common econometrics analyses are robust to the removal of a small proportion (e.g. one tenth of one percent of the data). Again using the empirical influence function, we provide an automated method and software package to answer this question.

We found that some common "gold-standard" applied econometrics papers can have central claims reversed by removing only a very small number of data points, even though there is no clear evidence of misspecification. A key theoretical takeaway of our work is that datsets with a low signal to noise ratio (defined as the effect size over the estimator standard deviation times $\sqrt{N}$) will always be sensitive to adversarial removal of a very small number of datapoints. Though the precise meaning will depend on the context, the broad implication is that studies which attempt to overcome low SNR with large sample sizes will be inherently non-robust, even in the absence of misspecification or gross errors.

2.3 Propagation of uncertainty in scalable Bayesian inference

Complex scientific inference procedures, such as the creation of astronomical catalogues, often exhibit uncertainty in many aspects of the model. For instance, in order to infer whether a handful of pixels on a telescopic image is a dim star or a distant galaxy, one must know the distortion (aka the point spread function) of the telescope, the lightness of the sky background, the noise of the photoreceptors, and the identity of nearby celestial objects, all of which quantities must themselves be inferred with some uncertainty.

Bayesian procedures coherently propagate uncertainty between all such model quantities, but classical MCMC procedures do not scale well, and are far beyond computational reach for astronomical catalogues. Researchers often turn to optimization-based mean field Variational Bayes (MFVB) procedures as a scalable alterative to MCMC, but MFVB does not estimate posterior correlations, and is known to underestimate marginal posterior uncertainties. [1]

In CITE, I develop a method to recover accurate posterior uncertainties from MFVB approximations without needing to fit a more complex model, or indeed to re-fit the original model. The idea is to exploit a duality between posterior covariances the sensitivity of posterior means and use the sensitivity of the MFVB approximation to infinitesimal perturbations as an estimator of the posterior covariance. We call the method "linear response variational Bayes" (LRVB) after the idea's progenitor as a method in statistical mechanics for inferring microscopic intensive thermodynamic quantities from macroscopic perturbations of extensive quantities. Computing the LRVB covaraince requires solving a linear system, which in scientific applications is often sparse and can be solved using iterative techniques such as conjugate gradient.

We compare LRVB covariances to MCMC on a large number of real-world datasets, including logistic regression on internet-scale data, the Cormack-Jolly-Seber model from ecology, and hierarchical generalized linear models from the social sciences, and demonstrated accurate posterior covariances computed over an order of magnitude faster than MCMC.

3 Selected Future work

There is a lot to be done simply applying the above methodology to applied problems in conjuction with collaborators with domain expertise. However, in this section, I will focus instead on new methodological directions suggested by the above work.

The bootstrap and the bootstrap after the bootstrap. Our work on the higher-order infinitesimal jackknife could be applied to other random reweighting schemes, particularly the bootstrap. The bootstrap is known to have frequentist properties that are asymptotically more accurate than the normal approximation in certain circumstances CITE, but the bootstrap requires re-computing an estimator as many time as there are bootstrap samples. However, a sufficiently high-order IJ estimate will approach the bootstrap estimator at a rate faster than the bootstrap's extra accuracy, strongly suggesting that the IJ will inherit all of the bootstrap's attractive properties at a fraction of the computational cost. The IJ is particularly appealing for bootstrap-after-bootstrap procedures, which have attractive theoretical properties but are computationally prohibitive even on medium-sized statistical problems. It seems plausible that the HOIJ could open up a range of bootstrap applications that are presently out of reach.

Bayesian model criticism. Essentially all tools for checking the accuracy of Bayesian models are frequentist in nature — e.g. checking whether the data is likely under draws from the prior or posterior, or evaluating its predicive

[1] The frequentist expectation-maximization, or EM, algorithm, can be understood as a MFVB procedure, and the present criticism applies to it as well.

performance on a held-out dataset. Predictive model checks, such as leave-one-out CV are attractive, but currently have few practical implementations that avoid multiple runs of the MCMC algorithm. However, the possibility of forming higher-order expansions of the Bayesian posterior as a function of the empirical distribtuion could change that.

Partitioned Bayesian inference with theoretical bounds. Given a large, complicated problem, it is often computationally convenient to perform inference in separate computatoinal steps, while still propagating uncertainty from one step to another. For example, in CITE, we first fit spline regressions to time series of gene expression data, and then clustered the spline fits to group together genes with similar behavior. The Bayesian ideal, which is the simultaneous estimation of the clusters and spline regression, was too computationally prohibitive, and would intuitively have given a similar result to the sequential analysis, to the extent that the cluster center priors did not shrink the spline regressions too much.

Incorporating LRVB corrections into MFVB approximations.

Bootstrapping simulation-based inference.